2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

# Thompson-Sampling Based Reinforcement Learning for Networked Control of Unknown Linear Systems

Borna Sayedana, Mohammad Afshari, Peter E. Caines, and Aditya Mahajan

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992565

*Abstract*— **In recent years, there has been considerable interest in reinforcement learning for linear quadratic Gaussian (LQG) systems. In this paper, we consider a generalization of such systems where the controller and the plant are connected over an unreliable packet drop channel. Packet drops cause the system dynamics to switch between controlled and uncontrolled modes. This switching phenomena introduces new challenges in designing learning algorithms. We identify a sufficient condition under which the regret of Thompson sampling-based reinforcement learning algorithm with dynamic episodes (TSDE) at horizon $T$ is bounded by $\tilde{\mathcal{O}}(\sqrt{T})$, where the $\tilde{\mathcal{O}}(\cdot)$ notation hides logarithmic factors in $T$. These are the first results to generalize regret bounds of LQG systems to packet-drop networked control models.**

## I. INTRODUCTION

Systems with linear dynamics, quadratic cost, and Gaussian noise, are a commonly used modelling framework in control theory. In recent years, there has been a significant interest in such LQG systems in the AI literature as well. Apart from the importance of LQG models for applications, a major reason for this interest is that LQG systems are the simplest models with continuous state and action spaces, and unbounded cost. Therefore, algorithms which learn to control unknown LQG systems must carefully design exploration schemes to ensure stability during learning.

A useful metric in analyzing the performance of an online learning algorithm is the notion of regret. Regret of an algorithm is defined to be the accumulated difference between the controller's policy and the optimal policy. It was recently shown in [1] that any learning algorithm for regular LQ must have a regret of $\tilde{\Omega}(n^{0.5}m\sqrt{T})$, where $n$ is the dimension of the state, $m$ is the dimension of the controls, $T$ is the horizon, and the $\tilde{\Omega}(\cdot)$ notation hides polylogarithmic factors in $T$. There are several algorithms [1]–[11] which nearly achieve this lower bound and have regret which can be upper bounded by $\tilde{\mathcal{O}}(n^{1.5}m\sqrt{T})$, where $\tilde{\mathcal{O}}(\cdot)$ notation hides polylogarithmic factors in $T$.

In this paper, we investigate a generalization of LQG models where the controller and the plant are connected over a noisy wireless channel. Such networked control systems (NCS) come up in various modern applications including platooning of self driving trucks and control of Internet of Things. There is a vast literature on planning for NCS [12], [13]. However, as far as we are aware, deriving regret bounds for learning in NCS has not been investigated in the literature.

A standard result in planning for NCS is that the networked control system can be stabilized if the "capacity" of the channel is greater than a quantity which depends on the unstable eigenvalue of the system. In the simple case of a packet drop channel, the corresponding condition is that the probability of packet drop must be less than $1/\lambda_{\max}^2$, where $\lambda_{\max}$ is the largest eigenvalue of the system. So the natural question in the context of learning is the following: what are the conditions on the packet drop probability to ensure that learning regret in NCS is $\tilde{\mathcal{O}}(\sqrt{T})$. In this paper, we provide an initial partial answer to this question.

We consider the simplest model of NCS where the controller and plant are connected over a packet-drop channel. When the channel is on, the plant receives the control action taken by the controller; however, when the channel is off, the control action is not received at the plant and the plant evolves in an open-loop manner. Thus the packet-drop nature of the channel introduces a non-linearity, which causes the dynamics to switch between closed-loop and open-loop behaviors. Thus, the analysis of existing algorithms is not directly applicable to NCS. There is some work on RL for NCS [14]–[16], but these papers do not characterize regret.

Due to their switching nature, NCS may be viewed as Switched Linear Systems (SLS) or Markov Jump Linear Systems (MJS), depending on the assumptions on packet drops. In recent years, there has been some work on adaptive control/reinforcement learning for MJS [17]. However, it was assumed in the model considered in [17] that the discrete state (or mode) at time $t$ is available to the controller when taking the action at time $t$. However, this is not the case for the NCS model that we consider in this paper. In our model, the controller doesn't know a priori if the control action is going to be dropped. So the result of [17] is not directly applicable to our model.

We consider Thompson Sampling with Dynamic Episode (TSDE) proposed by [10], which is a Bayesian algorithm for learning unknown LQ systems. We present the natural generalization of TSDE for NCS, and identify sufficient conditions under which the regret of TSDE is $\tilde{\mathcal{O}}(n^{1.5}m\sqrt{T})$. These conditions specify a relationship between the packet drop probability and the set of unknown parameters of the system. We present examples to show that these conditions on learning can be strictly stronger or same as the conditions for planning. This suggests that learning unknown NCS may require stronger conditions.

The rest of this paper is organized as follows. We introduce

The authors are with the Department of Electrical and Computer Engineering, McGill University, 3480 Rue University, Montreal, QC H3A 0E9, Canada. Emails: borna.sayedana@mail.mcgill.ca, mohammad.afshari2@mail.mcgill.ca, peterc@cim.mcgill.ca, aditya.mahajan@mcgill.ca.

978-1-6654-6761-2/22/$31.00 ©2022 IEEE

the model in Sec. II, describe the TSDE algorithm, assumptions, and our main result in Sec. III, discuss the salient features of the sufficient conditions in Sec. 4, and present key steps of the proof in Sec. V. Finally, we conclude in Sec. VI.

## II. Model and problem formulation

Consider a linear quadratic system with state $x_t \in \mathbb{R}^n$, control input $u_t \in \mathbb{R}^m$, and disturbance $w_t \in \mathbb{R}^n$. We assume that the system starts from an initial state $x_1 = 0$ and evolves over time according to

$$x_{t+1} = Ax_t + \nu_t B u_t + w_t, \quad t \geq 1, \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are the system dynamics matrices, the noise $\{w_t\}_{t\geq 1}$ is an independent and identically distributed Gaussian process with $w_t \sim \mathcal{N}(0, \sigma_w^2 I)$ and $\{\nu_t\}_{t\geq 1}$ is an i.i.d. Bernoulli process with $\mathbb{P}(\nu_t = 1) = q$.

At each time $t$, the system incurs a per-step cost given by

$$c(x_t, u_t, \nu_t) = x_t^\intercal Q x_t + \nu_t u_t^\intercal R u_t, \tag{2}$$

where $Q$ and $R$ are positive definite matrices.

Let $\theta^\intercal = [A, B]$ denote the parameters of the system. $\theta \in \mathbb{R}^{d\times n}$, where $d = n+m$. The performance of any policy $\pi = (\pi_1, \pi_2, \cdots)$ is measured by the long-term average cost given by

$$J(\pi; \theta) = \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}^\pi \Big[\sum_{t=1}^{T} c(x_t, u_t, \nu_t)\Big], \tag{3}$$

where the expectation is with respect to the prior on $\theta$, the noise processes, the channel processes, the initial conditions, and the potential randomizations done by the policy $\pi$.

Let $J(\theta)$ denote the infimum of $J(\pi; \theta)$ over all policies. Under the assumptions that pair $(A, B)$ is controllable and probability of successful transmission satisfies $1 - q \leq 1/|\lambda_{\max}(A)|^2$, it is shown in [18] that $J(\theta)$ is finite and is given by

$$J(\theta) = \sigma_w^2 \operatorname{Tr}(S(\theta)), \tag{4}$$

where $S(\theta)$ is the unique positive semi-definite solution of the following modified Riccati equation:

$$\begin{aligned} S(\theta) &= Q + A^\intercal S(\theta) A \\ &\quad - qA^\intercal S(\theta) B (R + B^\intercal S(\theta) B)^{-1} B^\intercal S(\theta) A. \end{aligned} \tag{5}$$

Furthermore, the optimal control action is given by

$$u_t = G(\theta) x_t, \tag{6}$$

where the gain matrix $G(\theta)$ is given by

$$G(\theta) = -(R + B^\intercal S(\theta) B)^{-1} B^\intercal S(\theta) A. \tag{7}$$

We are interested in the setting where the system parameters $(A, B)$ are unknown while the channel statistics $q$ and the cost matrices $(Q, R)$ are known. We denote the unknown parameters by a random variable $\theta$ and assume that there is a prior distribution on $\theta$. The Bayesian regret of a policy $\pi$ operating for horizon $T$ is defined by

$$\mathcal{R}(T; \pi) = \mathbb{E}^\pi \Big[\sum_{t=1}^{T} c(x_t, u_t, \nu_t) - TJ(\theta)\Big], \tag{8}$$

where the expectation is with respect to the prior on $\theta$, the noise processes, the channel processes, the initial conditions, and the potential randomizations done by the policy $\pi$.

## III. Thompson sampling based learning algorithm

### A. Prior and Posterior Beliefs

We assume that the unknown model parameters $\theta$ lie in a compact subset $\Theta$ of $\mathbb{R}^{d\times n}$. For any distribution $f$ on $\mathbb{R}^d$, we will use the notation $f\big|_\Theta$ to denote the projection of $f$ onto $\Theta$, i.e.,

$$f\big|_\Theta = \begin{cases} \frac{f(\theta)}{\int_\Theta f(\theta) d\theta} & \text{if} \quad \theta \in \Theta \\ 0 \quad \text{otherwise.} \end{cases}$$

For any $\theta \in \Theta$, let $\theta^k$ denote the $k$-th column of $\theta$ (thus, $\theta = [\theta^1, \cdots, \theta^n]$) and let $A_\theta$ and $B_\theta$ to denote the $A$ and $B$ matrices corresponding to $\theta$ (thus, $\theta^\intercal = [A_\theta, B_\theta]$).

We assume that $\theta$ is a random variable that is independent of the initial states, the noise processes, and the channel state process. Furthermore, we assume that there is a prior $p_1$ on $\theta$ that satisfies the following.

**Assumption 1** *$p_1$ is given as:*

$$p_1(\theta) = \Big[\prod_{i=1}^{n} \xi_1^i(\theta^i)\Big]\Big|_\Theta$$

*where for $i \in \{1, \cdots, n\}$, $\xi_1^i = \mathcal{N}(\mu_1^i, \Sigma_1)$ with mean $\mu_1^i \in \mathbb{R}^d$ and positive-definite covariance $\Sigma_1 \in \mathbb{R}^{d\times d}$.*

Let $z_t = \operatorname{vec}(x_t, \nu_t u_t)$. We can write the dynamics as

$$x_{t+1} = \theta^\intercal z_t + w_t. \tag{9}$$

We maintain a posterior distribution $\mu_t$ on $\theta$ based on the history $(x_{1:t-1}, u_{1:t-1}, \nu_{1:t-1})$ of the observations until time $t$. From standard results in Bayesian regression [19], we know that the posterior is a truncated Gaussian distribution

$$p_t(\theta) = \Big[\prod_{i=1}^{n} \xi_t^i(\theta^i)\Big]\Big|_\Theta$$

where for $i \in \{1, \cdots, n\}$, $\xi_t^i(\theta^i) = \mathcal{N}(\mu_t^i, \Sigma_t)$ and $\{\mu_t^i\}_{i=1}^n$ and $\Sigma_t$ can be updated recursively as follows:

$$\mu_{t+1}^i = \mu_t^i + \frac{\Sigma_t z_t (x_{t+1}^i - (\mu_t^i)^\intercal z_t)}{\sigma_w^2 + z_t^\intercal \Sigma_t z_t}, \tag{10}$$

$$\Sigma_{t+1}^{-1} = \Sigma_t^{-1} + \frac{1}{\sigma_w^2} z_t z_t^\intercal, \tag{11}$$

where $x_t = [x_t^1, \cdots, x_t^n]$.

**Algorithm 1** TSDE

```
input: $\Theta$, $\hat{\theta}$, $\Sigma_1$
initialization: $t \leftarrow 1$, $t_0 \leftarrow 0$, $k \leftarrow 0$.
for $t = 1, 2, \cdots$ do
    observe $x_t$
    update $p_t$ according to (10)–(11)
    if $\big((t - t_k > T_{k-1})$ or $(\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k})\big)$ then
        $T_k \leftarrow t - t_k$, $k \leftarrow k + 1$, $t_k \leftarrow t$
        sample $\theta_k \sim p_t$
    end if
    Apply control $u_t = G(\theta_k) x_t$
end for
```

### *B. Thompson Sampling with Dynamic Episodes Algorithm*

We now present a variation of the Thompson sampling with dynamic episodes (TSDE) algorithm of [10] for the networked control model presented in Sec. II.

As the name suggests, the algorithm operates in episodes of dynamic length. Let $t_k$ and $T_k$ denote the start time and the length of episode $k$, respectively. Episode $k$ ends when the length of the episode is strictly larger than the length of the previous episode (i.e., $t - t_k > T_{k-1}$) or at the first time after $t_k$ when the determinant of the covariance $\Sigma_t$ falls below half of its value at time $t_k$, i.e., $\det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k}$. Thus,

$$t_{k+1} = \min \left\{ t > t_k \,\middle|\, \begin{array}{l} t - t_k > T_{k-1} \text{ or} \\ \det \Sigma_t < \frac{1}{2} \det \Sigma_{t_k} \end{array} \right\}. \tag{12}$$

Note that the stopping condition (12) implies that

$$T_k \leq T_{k-1} + 1, \quad \forall k. \tag{13}$$

The TSDE algorithm works as follows. At the beginning of episode $k$, a parameter $\theta_k$ is sampled from the posterior distribution $p_{t_k}$. During the episode, the control inputs are generated using the sampled parameters $\theta_k$, i.e.,

$$u_t = G(\theta_k) x_t, \quad t_k \leq t \leq t_{k+1} - 1. \tag{14}$$

The complete algorithm is presented in Algorithm 1.

### *C. Regret Bounds*

We impose the following assumptions on the support of the prior distribution.

**Assumption 2** *For every $\theta \in \Theta$, the pair $(A_\theta, B_\theta)$ is controllable.*

**Assumption 3** *For all $\theta \in \Theta$, the probability of successful transmission satisfies the sufficient condition of [18]:*

$$1 - q \leq \frac{1}{|\lambda_{\max}(A_\theta)|^2}, \quad \forall \theta \in \Theta \tag{15}$$

*where $\lambda_{\max}(A_\theta)$ denotes the maximum eigenvalue of $A_\theta$.*

**Assumption 4** *Define $\delta$ and $\sigma$ as follows:*

$$\delta := \sup_{\theta, \phi \in \Theta} \|A_\theta + B_\theta G(\phi)\|,$$
$$\sigma := \sup_{\theta \in \Theta} \|A_\theta\|,$$

*where $\|A_\theta\|$ denotes the spectral norm of $A_\theta$. Then, we assume that the probability of successful transmission satisfies: $\delta^q \sigma^{1-q} < 1$ or equivalently:*

$$q \log(\delta) + (1 - q) \log(\sigma) < 0$$

*In [20] and [21], assumptions similar to Assumption 4 are imposed to ensure the almost sure stability of Markov jump systems (MJS).*

The following result provides an upper bound on the regret of the proposed algorithm.

**Theorem 1** *Under Assumptions 1–4, the regret of TSDE is upper bounded by*

$$\mathcal{R}(T; \texttt{TSDE}) \leq \tilde{\mathcal{O}}(\sigma_w^2 (n + m) \sqrt{nT}). \tag{16}$$

The proof is presented in Sec. V.

## IV. Discussion of the assumption

Assumptions 2 and 3 are necessary for the learning problem to be well posed. The additional technical assumption that we have is Assumption 4. Both Assumptions 3 and 4, pose a constraint between the packet drop probability $q$ and the uncertain set $\Theta$. In this section, we explore this relationship in details.

Define a feasible region for planning as, $\mathcal{Q}_p(\Theta) = [q_p, 1]$, where

$$q_p = \sup_{\theta \in \Theta} \Big[ 1 - \frac{1}{|\lambda_{\max}(A_\theta)|^2} \Big]^+,$$

where $[x]^+ = \max\{x, 0\}$. Similarly, define a feasible region for learning as, $\mathcal{Q}_\ell(\Theta) = \{q \in [0, 1] : \ \delta^q \sigma^{1-q} < 1\}$, where $\delta$ and $\sigma$ depend on $\Theta$ and are given in Assumption 4.

For the unknown system to have finite performance, $q \in \mathcal{Q}_p(\Theta)$. For our proof of the upper bound to hold $q \in \mathcal{Q}_p(\Theta) \cap \mathcal{Q}_\ell(\Theta)$. So, a natural question is whether $\mathcal{Q}_p(\Theta) \subset \mathcal{Q}_\ell(\Theta)$ or $\mathcal{Q}_\ell(\Theta) \subset \mathcal{Q}_p(\Theta)$. We consider four cases for $(\delta, \sigma)$ and answer this question for each case.

*a) Case 1: $\delta < 1$ and $\sigma < 1$.:* Observe that $\lambda_{\max}(A_\theta) \leq ||A_\theta|| \leq \sigma$. Therefore, if $\sigma < 1$,

$$\frac{1}{|\lambda_{\max}(A_\theta)|^2} \geq \frac{1}{\sigma^2} \geq 1.$$

Hence,

$$q_p = \sup_{\theta \in \Theta} \Big[ 1 - \frac{1}{|\lambda_{\max}(A_\theta)|^2} \Big]^+ = 0.$$

Thus, $\mathcal{Q}_p(\theta) = [0, 1]$.

Furthermore, $\delta < 1$ and $\sigma < 1$ implies that for all $q$, $\delta^q \sigma^{1-q} < 1$. Hence, $\mathcal{Q}_\ell(\Theta) = [0, 1]$.

Thus, in this case, both $\mathcal{Q}_\ell(\Theta) = \mathcal{Q}_p(\Theta) = [0, 1]$.

*b) Case 2: $\delta < 1$ and $\sigma > 1$.:* In this case, $\mathcal{Q}_\ell(\theta) = [q_\ell, 1]$, where

$$q_\ell = \frac{\log \sigma}{\log \sigma + \log \frac{1}{\delta}}.$$

We now show that depending on $\Theta$, $q_p > q_\ell$ or $q_p < q_\ell$.

**Example 1** *Suppose,*

$$\Theta = \{[A, B] \in \mathbb{R}^2 : A \in [0.8, 1.2] \text{ and } B \in [2.0, 2.4]\}.$$

*Then,* $\delta = 0.624$ *and* $\sigma = 1.2$*. Moreover,*

$$q_l = 0.279 \quad \textit{and} \quad q_p = 0.305.$$

*Thus* $q_p > q_\ell$ *and therefore,* $\mathcal{Q}_p(\Theta) \cap \mathcal{Q}_\ell(\Theta) = \mathcal{Q}_p(\Theta)$.

**Example 2** *Suppose,*

$$\Theta = \{[A, B] \in \mathbb{R}^2 : A \in [0.8, 1.2] \textit{ and } B \in [0.5, 0.9]\}.$$

*Then,* $\delta = 0.962$ *and* $\sigma = 1.2$*. Moreover,*

$$q_l = 0.824 \quad \textit{and} \quad q_p = 0.305.$$

*Thus* $q_\ell > q_p$ *and therefore,* $\mathcal{Q}_p(\Theta) \cap \mathcal{Q}_\ell(\Theta) = \mathcal{Q}_\ell(\Theta)$.

*c) Case 3:* $\delta > 1$ *and* $\sigma < 1$*.:* In this case $\mathcal{Q}_\ell(\Theta) = [0, q_\ell]$, where $q_\ell$ is same as Case 2, but can also be rewritten as

$$q_\ell = \frac{\log \frac{1}{\sigma}}{\log \frac{1}{\sigma} + \log \delta}.$$

As in Case 1, $\sigma < 1$ implies that $\mathcal{Q}_p(\Theta) = [0, 1]$. Now, we present an example to show that $q_\ell$ can be less than 1.

**Example 3** *Suppose,*

$$\Theta = \{[A, B] \in \mathbb{R}^2 : A \in [0.1, 0.9] \textit{ and } B \in [1.0, 2.4]\}.$$

*Then,* $\delta = 1.190$ *and* $\sigma = 0.9$*. Moreover,*

$$q_\ell = 0.37 \quad \textit{and} \quad q_p = 0.$$

*Thus* $q_\ell > q_p$ *and therefore,* $\mathcal{Q}_p(\Theta) \cap \mathcal{Q}_\ell(\Theta) = \mathcal{Q}_\ell(\Theta)$.

*d) Case 4:* $\delta > 1$ *and* $\sigma > 1$*.:* In this case, $\delta^q \sigma^{1-q} > 1$ and hence, $\mathcal{Q}_\ell(\Theta) = \emptyset$.

The above examples show that in some instances, $\mathcal{Q}_p(\Theta) \subset \mathcal{Q}_\ell(\Theta)$, while in others $\mathcal{Q}_p(\Theta) \supset \mathcal{Q}_\ell(\Theta)$.

We conjecture that, Assumption 4 is stronger than what it needs to be and it should be possible to relax it and replace $\|\cdot\|$ in the definition of $\delta$ and $\sigma$ by the spectral radius of the respective matrices. This would require modifying the proof of Lemma 1 in Sec. V to exploit asymptotic stability of $A + \nu_k B G_k$ rather than the contractive property. We refer the reader to [22], where a similar relaxation for the original TSDE algorithm of [10] is presented.

## V. Regret analysis

*a) A preliminary result.:* We first start with a preliminary result, which is critical in deriving the regret bounds.

**Lemma 1** *Define* $\gamma_t = \delta^{\nu_t} \sigma^{1-\nu_t}$ *and for any* $s \le t + 1$*, define* $\Gamma_{s,t} = \gamma_s \cdots \gamma_t$*. Then, under* Assumption 4*, there exists a* $\bar{\Gamma} < \infty$ *such that for all* $t > 1$*,*

$$\sum_{s=1}^{t-1} \Gamma_{s+1,t-1} \le \bar{\Gamma}, \quad \textit{a.s.}$$

*Proof:* We use the result established in [21] for Markov jump systems (MJS) to prove this lemma. Consider a switched linear system with two modes, $A_1 = A_\theta$, $A_2 = A_\theta + B_\theta G(\phi)$ and i.i.d. probability of transition $p = (1-q, q)$. Then Assumption 4 implies [21, Assumption 2] and the result follows from [21, Lemma 1]. ■

*b) Regret decomposition.:* For the ease of notation, we use $\mathcal{R}(T)$ instead of $\mathcal{R}(T; \texttt{TSDE})$ in this section. We also use $G_k$ and $S_k$ to denote $G(\theta_k)$ and $S(\theta_k)$ respectively. We know that the policy $u_t = G_k x_t$ is optimal for model $\theta_k$ and, therefore, satisfies the following Bellman equation:

$$\begin{aligned} J(\theta_k) + x_t^\intercal S_k x_t &= c(x_t, u_t, \nu_t) \\ &+ \mathbb{E}[(\theta_k^\intercal z_t + w_t)^\intercal S_k (\theta_k^\intercal z_t + w_t)]. \end{aligned} \tag{17}$$

Note that $x_{t+1} = \theta^\intercal z_t + w_t$. Adding and subtracting $\mathbb{E}[x_{t+1}^\intercal S_k x_{t+1}]$ in (17) and rearranging terms, we get

$$\begin{aligned} c(x_t, u_t, \nu_t) &= J(\theta_k) + x_t^\intercal S_k x_t - \mathbb{E}[x_{t+1}^\intercal S_k x_{t+1}] \\ &+ \mathbb{E}\big[(\theta^\intercal z_t)^\intercal S_k (\theta^\intercal z_t) - (\theta_k^\intercal z_t)^\intercal S_k (\theta_k^\intercal z_t)\big]. \end{aligned} \tag{18}$$

Let $K_T$ denote the number of episodes until horizon $T$. For each $k > K_T$, we define $t_k$ be to $T+1$. Then, using (18), we have that

$$\begin{aligned} \mathcal{R}(T) = &\underbrace{\mathbb{E}\Big[\sum_{k=1}^{K_T} T_k J(\theta_k) - T J(\theta)\Big]}_{\text{regret due to sampling error}=:\mathcal{R}_0(T)} \\ &+ \underbrace{\mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \big[x_t^\intercal S_k x_t - x_{t+1}^\intercal S_k x_{t+1}\big]\Big]}_{\text{regret due to time-varying controller}=:\mathcal{R}_1(T)} \\ &+ \underbrace{\mathbb{E}\Big[\sum_{k=1}^{K_T} \sum_{t=t_k}^{t_{k+1}-1} \big[(\theta^\intercal z_t)^\intercal S_k (\theta^\intercal z_t) - (\theta_k^\intercal z_t)^\intercal S_k (\theta_k^\intercal z_t)\big]\Big]}_{\text{regret due to model mismatch}=:\mathcal{R}_2(T)}. \end{aligned} \tag{19}$$

Thus,

$$\mathcal{R}(T) = \mathcal{R}_0(T) + \mathcal{R}_1(T) + \mathcal{R}_2(T). \tag{20}$$

We establish the bound on $\mathcal{R}(T)$ by individually bounding $\mathcal{R}_0(T)$, $\mathcal{R}_1(T)$, and $\mathcal{R}_2(T)$.

*c) Bound on individual terms of* (20)*.:* Let $X_T = \sigma_w + \max_{t \le T} \|x_t\|$ be the maximum value of the norm of the state. Recall that $K_T$ is the number of episodes until horizon $T$. Then, we have the following.

**Proposition 1** *The terms in* (20) *are bounded as follows:*

1) $\mathcal{R}_0(T) \le \mathcal{O}(\sigma_w^2 \mathbb{E}[K_T])$.
2) $\mathcal{R}_1(T) \le \mathcal{O}(\mathbb{E}[K_T X_T^2])$.
3) $\mathcal{R}_2(T) \le \mathcal{O}\big(d\sqrt{nT}\, \mathbb{E}[\sqrt{(\sigma_w^2 X_T^2 + X_T^4) \log(T X_T^2)}]\big)$.

The proof of bounding $\mathcal{R}_0(T)$ and $\mathcal{R}_1(T)$ are similar to those in [10] and is omitted. The proof argument for bounding $\mathcal{R}_2(T)$ is presented in Appendix VI-A.

*d) Bounding* $X_T$ *and* $K_T$*.:* Now, to prove the regret bounds, we establish the following bounds on $X_T$ and $K_T$.

**Lemma 2** *For any* $\alpha \ge 1$*, the following inequalities hold:*

1) $\mathbb{E}[X_T^\alpha] \le \mathcal{O}(\sigma_w^\alpha (\log T)^{\alpha/2})$.
2) $\mathbb{E}[X_T^\alpha \log X_T^2] \le \sigma_w^\alpha \tilde{\mathcal{O}}(1)$.
3) $K_T \le \mathcal{O}(\sqrt{dT \log(T X_T^2 / \sigma_w^2 d)})$.

We prove the bound on $\mathbb{E}[X_T^\alpha]$ below. The bounds on the other two terms can be proved in a manner similar to [10] and [22].

*Proof:* [of Part 1] During the $k$-th episode, we have $u_t = \nu_t G_k x_t$. Therefore,

$$\begin{aligned}\|x_{t+1}\| &= \|(A + \nu_t B G_k) x_t + w_t\| \\ &\le \|(A + \nu_t B G_k)\| \|x_t\| + \|w_t\| \\ &\le \gamma_t \|x_t\| + \|w_t\|, \end{aligned} \tag{21}$$

where the last inequality follows from the definition of $\gamma_t = \delta^{\nu_t} \sigma^{1-\nu_t}$. Then, iteratively applying (21), we get

$$\|x_t\| \le \sum_{s=1}^{t-1} \Gamma_{s+1,t-1} \|w_s\| \le \sum_{s=1}^{t-1} \Gamma_{s+1,t-1} \max_{s \le T} \|w_s\|, \tag{22}$$

where $\Gamma_{s+1,t-1} = \gamma_s \cdots \gamma_{t-1}$. Now using Lemma 1 in (22), we get

$$\|x_t\| \le \bar{\Gamma} \max_{s \le T} \|w_s\| \quad \text{a.s.} \tag{23}$$

Therefore, for any $\alpha \ge 1$,

$$\begin{aligned} X_T^\alpha &\le \big(\sigma_w + \bar{\Gamma} \max_{t \le T} \|w_t\|\big)^\alpha \\ &= \sum_{\ell=0}^{\alpha} \big(\alpha \ell\big)\, \sigma_w^{\alpha-\ell} (\bar{\Gamma} \max_{t \le T} \|w_t\|)^\ell \quad \text{a.s.} \end{aligned} \tag{24}$$

From [23] (or equivalently, see [24, Lemma 9] for a detailed proof), we have

$$\mathbb{E}\Big[\max_{t \le T} \|w_t\|^\ell\Big] \le \mathcal{O}(\sigma_w^\ell (\log T)^{\ell/2}).$$

Substituting this in (24), we get the result. ■

*e) Putting everything together:* An immediate consequence of Proposition 1 and Lemma 2 is the following.

**Corollary 1** *The terms in* (20) *are bounded as follows:*

1) $\mathcal{R}_0(T) \le \tilde{\mathcal{O}}(\sigma_w^2 \sqrt{dT})$.
2) $\mathcal{R}_1(T) \le \tilde{\mathcal{O}}(\sigma_w^2 \sqrt{dT})$.
3) $\mathcal{R}_2(T) \le \tilde{\mathcal{O}}\big(\sigma_w^2 d \sqrt{nT}\big)$.

*Proof:* We prove each part separately.

1) We have that

$$\begin{aligned}\mathcal{R}_0(T) &\le \mathcal{O}(\sigma_w^2 \mathbb{E}[K_T]) \\ &\overset{(a)}{\le} \mathcal{O}(\sigma_w^2 \mathbb{E}[\sqrt{dT \log(T X_T^2 / \sigma_w^2 d)}) \\ &\overset{(b)}{\le} \mathcal{O}(\sigma_w^2 \sqrt{dT} \log(\mathbb{E}[X_T^2] T / \sigma_w^2 d)) \\ &\overset{(c)}{\le} \mathcal{O}(\sigma_w^2 \sqrt{dT} \log(T/d)) \\ &\le \tilde{\mathcal{O}}(\sigma_w^2 \sqrt{dT}) \end{aligned} \tag{25}$$

where $(a)$ and $(c)$ follow from Lemma 2, and $(b)$ follows from Jensen's inequality.

2) We have that

$$\begin{aligned}\mathcal{R}_1(T) &\le \mathcal{O}(\mathbb{E}[K_T X_T^2]) \\ &\overset{(d)}{\le} \mathcal{O}(\mathbb{E}[X_T^2 \sqrt{dT \log(T X_T^2/\sigma_w^2 d)}] \\ &\le \mathcal{O}(\sqrt{dT} \mathbb{E}[X_T^2 \sqrt{\log(T X_T^2/\sigma_w^2 d)}]) \\ &\overset{(e)}{\le} \mathcal{O}(\sqrt{dT} \sqrt{\mathbb{E}[X_T^4 \log(T X_T^2/\sigma_w^2 d)]}) \\ &\overset{(f)}{\le} \mathcal{O}(\sqrt{dT} \sqrt{\sigma_w^4 \tilde{\mathcal{O}}(1)}) \\ &\le \tilde{\mathcal{O}}(\sigma_w^2 \sqrt{dT}), \end{aligned} \tag{26}$$

where $(d)$ follows from Lemma 2, $(e)$ follows from Jensens's inequality, $(f)$ follows from Lemma 2.

3) Observe that

$$\begin{aligned} &\mathbb{E}[\sqrt{(\sigma_w^2 X_T^2 + X_T^4) \log(T X_T^2)}] \\ &= \sigma_w^2 \mathbb{E}\Bigg[\sqrt{\bigg(\frac{X_T^2}{\sigma_w^2} + \frac{X_T^4}{\sigma_w^4}\bigg) \log\bigg(T \sigma_w^2 \frac{X_T^2}{\sigma_w^2}\bigg)}\Bigg] \\ &\le \sigma_w^2 \tilde{\mathcal{O}}(1), \end{aligned} \tag{27}$$

where the last inequality follows similarly to the argument in (25). Substituting the value of (27) in the expression of $\mathcal{R}_2(T)$ in Proposition 1 gives us the result.

■

*f) Proof of Theorem 1:* Now we have all the ingredients to prove Theorem 1.

*Proof:* [of Theorem 1] Corollary 1 implies that the $\mathcal{R}_2(T)$ term dominates $\mathcal{R}_0(T)$ and $\mathcal{R}_1(T)$. Thus, the total regret is of the same order as $\mathcal{R}_2(T)$. ■

## VI. Conclusion

In this paper, we considered the problem of learning the optimal control policy in a networked control system where the link between the controller and the system is a packet drop channel. We identified sufficient conditions under which the regret of TSDE is bounded by $\tilde{\mathcal{O}}(n^{1.5} m \sqrt{T})$. This bound is same as the regret bound for classical LQG systems. Our results show that, as long as the packet-drop probability satisfy specific conditions that depend on the set $\Theta$ of uncertain parameters, learning for NCS has a similar regret as classical LQG systems.

Our sufficient conditions are the intersection of two feasible regions for the packet drop probability: the feasible region $\mathcal{Q}_p(\Theta)$ for planning and the feasible region $\mathcal{Q}_\ell(\Theta)$ for learning. We present examples to show that none of these two conditions are more restrictive than the other one. These conditions are sufficient conditions and motivate further investigation into the model, in particular, to identify lower bounds on the regret and investigating more sophisticated models of networked control systems.

## APPENDIX

### A. Proof of Proposition 1-Bound on $\mathcal{R}_2(T)$

We start by considering the term inside the expectation of $\mathcal{R}_2(T)$:

$$\begin{aligned}
&\|S_k^{0.5}\theta^\intercal z_t\|^2 - \|S_k^{0.5}\theta_k^\intercal z_t\|^2\\
&= \big(\|S_k^{0.5}\theta^\intercal z_t\| + \|S_k^{0.5}\theta_k^\intercal z_t\|\big)\big(\|S_k^{0.5}\theta^\intercal z_t\| - \|S_k^{0.5}\theta_k^\intercal z_t\|\big)\\
&\le (\|S_k^{0.5}\theta^\intercal z_t\| + \|S_k^{0.5}\theta_k^\intercal z_t\|)\|S_k^{0.5}(\theta-\theta_k)^\intercal z_t\|\\
&\le (\|S_k^{0.5}\theta^\intercal z_t\| + \|S_k^{0.5}\theta_k^\intercal z_t\|)\|S_k^{0.5}\|\|(\theta-\theta_k)^\intercal z_t\|. \qquad (28)
\end{aligned}$$

Note that we can bound $\|S_k^{0.5}\theta^\intercal z_t\|\|S_k^{0.5}\| \le \|S_k^{0.5}\|\|\theta^\intercal\|\|[I,\nu_t G_k^\intercal]\|\|x_t\|\|S_k^{0.5}\| \le \mathcal{O}(\|x_t\|)$ because each of the other terms are bounded as $\theta$ and $\theta_k$ belong to a compact set. By the same argument $\|S_k^{0.5}\theta_k^\intercal z_t\| \le \mathcal{O}(\|x_t\|)$. Combining this with the fact that $\|x_t\| \le X_T$ and substituting in (28), we get

$$\|S_k^{0.5}\theta^\intercal z_t\|^2 - \|S_k^{0.5}\theta_k^\intercal z_t\|^2 \le \mathcal{O}\big(X_T\|(\theta-\theta_k)^\intercal z_t\|\big). \quad (29)$$

Therefore,

$$\mathcal{R}_2(T) \le \mathcal{O}\Bigg(\mathbb{E}\Bigg[X_T\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta-\theta_k)^\intercal z_t\|\Bigg]\Bigg). \quad (30)$$

Now, we consider the term inside the $\mathcal{O}(\cdot)$:

$$\begin{aligned}
&\mathbb{E}\Bigg[X_T\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta-\theta_k)^\intercal z_t\|\Bigg]\\
&= \mathbb{E}\Bigg[X_T\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta-\theta_k)^\intercal \Sigma_{t_k}^{-0.5}\Sigma_{t_k}^{0.5}z_t\|\Bigg]\\
&\le \mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta-\theta_k)^\intercal \Sigma_{t_k}^{-0.5}\| \times X_T\|\Sigma_{t_k}^{0.5}z_t\|\Bigg]\\
&\le \sqrt{\mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta-\theta_k)^\intercal \Sigma_{t_k}^{-0.5}\|^2\Bigg]}\\
&\quad\times\sqrt{\mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}X_T^2\|\Sigma_{t_k}^{0.5}z_t\|^2\Bigg]}, \qquad (31)
\end{aligned}$$

where the last inequality follows from Cauchy-Schwartz inequality.

Now, we bound the two terms in (31) separately in Lemmas 3 and 4.

**Lemma 3** *We have the following inequality*

$$\mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}\|(\theta-\theta_k)^\intercal \Sigma_{t_k}^{-0.5}\|^2\Bigg] \le \mathcal{O}(dnT).$$

*Proof:* The proof is similar to the proof of [22, Lemma 7]. ■

**Lemma 4** *We have the following inequality*

$$\mathbb{E}\Bigg[\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}X_T^2\|\Sigma_{t_k}^{0.5}z_t\|^2\Bigg] \le \mathcal{O}\big(d(\sigma_w^2X_T^2+X_T^4)\log(TX_T^2)\big).$$

*Proof:* The proof follows a similar structure as the proof of [22, Lemma 8].

For any $s \le t$, Eq. (11) implies that $\Sigma_s^{-1} \preceq \Sigma_t^{-1}$ and consequently $\Sigma_s^{-1} \preceq \Sigma_t^{-1}$ implies that $\Sigma_s \succeq \Sigma_t$. Therefore, from [2, Lemma 11], we get that for any $V \neq 0$ (of appropriate dimensions),

$$\frac{\|V^\intercal\Sigma_sV\|}{\|V^\intercal\Sigma_tV\|} \le \frac{\det\Sigma_s}{\det\Sigma_t} = \frac{\det\Sigma_t^{-1}}{\det\Sigma_s^{-1}}. \quad (32)$$

Eq. (32) implies that for any $t \in \{t_k,\cdots,t_{k+1}-1\}$, we have

$$\|\Sigma_{t_k}^{0.5}z_t\|^2 = z_t^\intercal\Sigma_{t_k}z_t \le \frac{\det\Sigma_t^{-1}}{\det\Sigma_{t_k}^{-1}}z_t^\intercal\Sigma_tz_t \le 2z_t^\intercal\Sigma_tz_t \quad (33)$$

where the last inequality follows from the second stopping criterion. Therefore,

$$\sum_{k=1}^{K_T}\sum_{t=t_k}^{t_{k+1}-1}X_T^2\|\Sigma_{t_k}^{0.5}z_t\|^2 \le 2X_T^2\sum_{t=1}^{T}z_t^\intercal\Sigma_tz_t. \quad (34)$$

Since $\Sigma_t \preceq \Sigma_1$, we have $\lambda_{\max}(\Sigma_t) \le \lambda_{\max}(\Sigma_1) = 1/\lambda_{\min}(\Sigma_1^{-1})$. Therefore for any $t$

$$z_t^\intercal\Sigma_tz_t \le \frac{1}{\lambda_{\min}(\Sigma_1^{-1})}\|z_t\|^2 \le \frac{1}{\lambda_{\min}(\Sigma_1^{-1})}M_G^2X_T^2, \quad (35)$$

where $M_G = \sup_{\delta\in\{0,1\},\theta\in\Theta}\|[I,\delta G(\theta)^\intercal]^\intercal\|$. From (35), we get that

$$z_t^\intercal\Sigma_tz_t \le \max\Bigg(\sigma_w^2, \frac{M_G^2X_T^2}{\lambda_{\min}(\Sigma_1^{-1})}\Bigg)\min\Bigg(1, \frac{z_t^\intercal\Sigma_tz_t}{\sigma_w^2}\Bigg). \quad (36)$$

Hence

$$\sum_{t=1}^{T}z_t^\intercal\Sigma_tz_t \le \Bigg(\sigma_w^2 + \frac{M_G^2X_T^2}{\lambda_{\min}(\Sigma_1^{-1})}\Bigg)\sum_{t=1}^{T}\min\Bigg(1, \frac{z_t^\intercal\Sigma_tz_t}{\sigma_w^2}\Bigg) \quad (37)$$

Using (11) and the intermediate step of the proof of [25, Lemma 6], we have

$$\begin{aligned}
\sum_{t=1}^{T}\min\Bigg(1, \frac{z_t^\intercal\Sigma_tz_t}{\sigma_w^2}\Bigg) &= \sum_{t=1}^{T}\min\Bigg(1, \Bigg\|\frac{\Sigma_t^{0.5}z_tz_t^\intercal\Sigma_t^{0.5}}{\sigma_w^2}\Bigg\|\Bigg)\\
&\le 2d\log\Bigg(\frac{\mathrm{Tr}(\Sigma_{T+1}^{-1})}{d}\Bigg) - \log\det\Sigma_1^{-1}. \quad (38)
\end{aligned}$$

Now, from (11), we get that

$$\begin{aligned}
\mathrm{Tr}(\Sigma_{T+1}^{-1}) &= \mathrm{Tr}(\Sigma_1^{-1}) + \sum_{t=1}^{T}\frac{1}{\sigma_w^2}\mathrm{Tr}(z_tz_t^\intercal)\\
&\le \mathrm{Tr}(\Sigma_1^{-1}) + \frac{T}{\sigma_w^2}M_G^2X_T^2, \quad (39)
\end{aligned}$$

where the last inequality uses the fact that $\mathrm{Tr}(z_tz_t^\intercal) = \mathrm{Tr}(z_t^\intercal z_t) = \|z_t\|^2 \le M_G^2X_T^2$. Combining (37) with (38) and (39), we get

$$\sum_{t=1}^{T}z_t^\intercal\Sigma_tz_t \le \mathcal{O}\big(d(\sigma_w^2+X_T^2)\log(TX_T^2)\big). \quad (40)$$

Therefore, we can bound the expectation of the right hand side of (34) as

$$\mathbb{E}\left[2X_T^2 \sum_{t=1}^{T} z_t^\intercal \Sigma_t z_t\right] \le \mathcal{O}\big(d(\sigma_w^2 X_T^2 + X_T^4)\log(TX_T^2)\big). \tag{41}$$

The result then follows from (34) and (41). ■

The proof for Proposition 1 then completes by substituting the results of Lemma 3 and 4 in (31).

### B. Proof of Lemma 2

We prove the last two parts separately.

**Lemma 5** *For any $\alpha \ge 1$, we have*

$$\mathbb{E}[X_T^\alpha \log X_T^2] \le \sigma_w^\alpha \tilde{\mathcal{O}}(1).$$

*Proof:* The proof argument is the same as that of [22, Lemma 5]. ■

**Lemma 6** *The number of episodes $K_T$ is bounded by*

$$K_T \le \mathcal{O}(\sqrt{dT\log(TX_T^2/\sigma_w^2 d)}).$$

*Proof:* The proof follows along the same lines as proof of [10, Lemma 3].

Define macro episodes with start times $t_{n_i}$, $i \in \mathbb{N}_{>0}$, where $n_1 = 1$ and for $i \ge 1$,

$$n_{i+1} = \min\left\{k > n_i \,\middle|\, \det\Sigma_{t_k} < \tfrac{1}{2}\det\Sigma_{t_{k-1}}\right\}.$$

Thus, a new macro-episode starts whenever an episode ends due to the second stopping criterion. Let $M$ denote the number of macro-episodes until time $T$ and define $n_{M+1} = K_T + 1$. Let $\tilde{T}_i$ denote the length of the $i$-th macro-episode. Within a macro-episode, all but the last episode must be triggered by the first stopping criterion. Thus, for $k \in \{n_i, n_i + 1, \cdots, n_{i+1} - 2\}$, $T_k = T_{k-1} + 1$. Hence, $T_k \ge (k - n_i + 1)$. Hence,

$$\begin{aligned}\tilde{T}_i &= \left[\sum_{k=n_i}^{n_{i+1}-2} T_k\right] + T_{n_{i+1}-1} \\ &\ge \sum_{j=1}^{n_{i+1}-n_i-1} (j+1) + 1 \ge \tfrac{1}{2}(n_{i+1}-n_i)^2 \end{aligned} \tag{42}$$

Hence,

$$n_{i+1} - n_i \le \sqrt{2\tilde{T}_i}, \quad \forall i \in \{1, \cdots, M\}. \tag{43}$$

Now, we know that

$$\begin{aligned}K_T = n_{M+1} - 1 &= \sum_{i=1}^{M}(n_{i+1} - n_i) \overset{(a)}{\le} \sum_{i=1}^{M}\sqrt{2\tilde{T}_i} \\ &\overset{(b)}{\le} \sqrt{M\sum_{i=1}^{M} 2\tilde{T}_i} = \sqrt{2MT}\end{aligned} \tag{44}$$

where $(a)$ uses (43) and $(b)$ uses the Cauchy-Schwartz inequality.

Now, observe that

$$\begin{aligned}\det\Sigma_T^{-1} &\overset{(c)}{\ge} \det\Sigma_{t_{n_M}}^{-1} \overset{(d)}{\ge} 2\det\Sigma_{t_{n_{M-1}}}^{-1} \\ &\ge \cdots \ge 2^{M-1}\det\Sigma_1^{-1},\end{aligned} \tag{45}$$

where $(c)$ follows because $\{\det\Sigma_t^{-1}\}_{t\ge 1}$ is a non-decreasing sequence (because $\Sigma_1^{-1} \preceq \Sigma_2^{-1} \ldots$) and $(d)$ and subsequent inequalities follow from the definition of the macro episode and the second triggering condition.

Since $\mathrm{Tr}(\Sigma_T^{-1}/d) \ge (\det\Sigma_T^{-1})^{1/d}$, we have

$$\begin{aligned}\mathrm{Tr}(\Sigma_T^{-1}) &\ge d(\det\Sigma_T^{-1})^{1/d} \overset{(e)}{\ge} d2^{(M-1)/d}(\det\Sigma_1^{-1})^{1/d} \\ &\ge d2^{(M-1)/d}\lambda_{\min}(\Sigma_1^{-1}),\end{aligned}$$

where $(e)$ comes from (45). Hence,

$$M \le 1 + d\log\frac{\mathrm{Tr}(\Sigma_T^{-1})}{d\lambda_{\min}(\Sigma_1^{-1})} \tag{46}$$

From (11), we know that

$$\Sigma_T^{-1} = \Sigma_1^{-1} + \frac{1}{\sigma_w^2}\sum_{t=1}^{T-1} z_t z_t^\intercal$$

Therefore,

$$\mathrm{Tr}(\Sigma_T^{-1}) = \mathrm{Tr}(\Sigma_1^{-1}) + \frac{1}{\sigma_w^2}\sum_{t=1}^{T-1} z_t^\intercal z_t \le \mathcal{O}(TX_T^2/\sigma_w^2) \tag{47}$$

where the last inequality uses the fact that $\|z_t\| = \|[I, \nu_t G_k^\intercal]^\intercal x_t\| \le \mathcal{O}(\|x_t\|)$ because $\theta_k$ belongs to a compact set and, by definition, $\|x_t\| \le X_T$.

Substituting (47) in (46), we get

$$M \le \mathcal{O}(d\log(TX_T^2/\sigma_w^2 d)).$$

Combining this with (44), we get the result. ■

## References

[1] M. Simchowitz and D. Foster, "Naive exploration is optimal for online lqr," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8937–8948.

[2] Y. Abbasi-Yadkori and C. Szepesvári, "Regret bounds for the adaptive control of linear quadratic systems," in *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 2011, pp. 1–26.

[3] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Optimism-based adaptive regulation of linear-quadratic systems," *IEEE Trans. Autom. Control*, vol. 66, no. 4, pp. 1802–1808, 2021.

[4] A. Cohen, T. Koren, and Y. Mansour, "Learning linear-quadratic regulators efficiently with only $\sqrt{T}$ regret," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1300–1309.

[5] M. Abeille and A. Lazaric, "Efficient optimistic exploration in linear-quadratic regulators via lagrangian relaxation," *ArXiv*, vol. abs/2007.06482, 2020.

[6] S. Dean, H. Mania, N. Matni, B. Recht, and S. Tu, "Regret bounds for robust adaptive control of the linear quadratic regulator," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 4192–4201.

[7] H. Mania, S. Tu, and B. Recht, "Certainty equivalent control of LQR is efficient," 2019, arXiv:1902.07826.

[8] M. K. S. Faradonbeh, A. Tewari, and G. Michailidis, "Input perturbations for adaptive control and learning," *Automatica*, vol. 117, p. 108950, 2020.

[9] ——, "On adaptive Linear–Quadratic regulators," *Automatica*, vol. 117, p. 108982, Jul. 2020.

[10] Y. Ouyang, M. Gagrani, and R. Jain, "Posterior sampling-based reinforcement learning for control of unknown linear systems," *IEEE Trans. Autom. Control*, vol. 65, no. 6, pp. 3600–3607, 2020.

[11] M. Abeille and A. Lazaric, "Improved regret bounds for Thompson sampling in linear quadratic control problems," in *International Conference on Machine Learning*, 2018, pp. 1–9.

[12] P. J. Antsaklis and J. Baillieul, Eds., *IEEE Trans. Autom. Control: Special Issue on Networked Control Systems*, vol. 49, no. 9, Sep. 2004.

[13] ——, *Proc. IEEE: Specical issue on Technology of Networked Control Systems*, vol. 95, no. 1, Jan. 2007.

[14] Y. Jiang, J. Fan, T. Chai, F. L. Lewis, and J. Li, "Tracking control for linear discrete-time networked control systems with unknown dynamics and dropout," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 10, pp. 4607–4620, 2017.

[15] J. Fan, Q. Wu, Y. Jiang, T. Chai, and F. L. Lewis, "Model-free optimal output regulation for linear discrete-time lossy networked control systems," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 50, no. 11, pp. 4033–4042, 2019.

[16] J. Li, Z. Xiao, P. Li, and Z. Ding, "Networked controller and observer design of discrete-time systems with inaccurate model parameters," *ISA transactions*, vol. 98, pp. 75–86, 2020.

[17] Y. Sattar, Z. Du, D. A. Tarzanagh, N. Ozay, L. Balzano, and S. Oymak, "Identification and adaptive control of markov jump systems: Sample complexity and regret bounds," in *ICML Workshop on Reinforcement Learning Theory*, 2021.

[18] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, and S. Sastry, "An LQG optimal linear controller for control systems with packet losses," in *Proceedings of the 44th IEEE Conference on Decision and Control*. IEEE, 2005, pp. 458–463.

[19] J. Sternby, "On consistency for the method of least squares using martingale theory," *IEEE Transactions on Automatic Control*, vol. 22, no. 3, pp. 346–352, 1977.

[20] Y. Fang, K. A. Loparo, and X. Feng, "Almost sure and $\delta$moment stability of jump linear systems," *International Journal of Control*, vol. 59, no. 5, pp. 1281–1307, 1994.

[21] B. Sayedana, M. Afshari, P. E. Caines, and A. Mahajan, "Consistency and rate of convergence of switched least squares system identification for autonomous switched linear systems," *arXiv preprint arXiv:2112.10753*, 2021.

[22] M. Gagrani, S. Sudhakara, A. Mahajan, A. Nayyar, and Y. Ouyang, "A relaxed technical assumption for posterior sampling-based reinforcement learning for control of unknown linear systems," *arXiv preprint arXiv:2108.08502*, 2021.

[23] P. J. Downey and R. S. Maier, "Stochastic orderings and the growth of expected extremes," Tech. Report 90-9, Department of Computer Science, University of Arizona, Tech. Rep., 1990.

[24] A. M. A. N. Mukul Gagrani, Sagar Sudhakara and Y. Ouyang, "A modified thompson sampling-based learning algorithm for unknown linear systems," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022.

[25] Y. Abbasi-Yadkori and C. Szepesvari, "Bayesian optimal control of smoothly parameterized systems: The lazy posterior sampling algorithm," 2014, arXiv preprint arXiv:1406.3926.